Article:
Fan, Z., Huang, Z., Chen, Z. et al. (3 more authors) (2024) Lightweight multiperson pose estimation with staggered alignment self-distillation. IEEE Transactions on Multimedia, 26. pp. 9228-9240. ISSN 1520-9210

https://doi.org/10.1109/TMM.2024.3387754

Lightweight Multiperson Pose Estimation with Staggered Alignment Self-distillation

Zhenkun Fan, Zhuoxu Huang, Zhixiang Chen, Tao Xu, Jungong Han, *Senior Member, IEEE,* and Josef Kittler *Fellow, IEEE*

***Abstract*—Accurate 2D human pose estimation from images is vital for understanding human actions. However, deploying the latest models, e.g., regression-based models, on resource-limited devices remains challenging due to their high computational requirements. In this paper, we address the resolution dilemma in regression-based multiperson pose estimation, where low-resolution inputs cause performance degradation, while high-resolution inputs drastically increase computational costs. To achieve a lightweight regression approach, it becomes crucial to enhance the model's capabilities in low-resolution scenarios. We propose the staggered alignment self-distillation (SASD) method and a corresponding network architecture. Our approach involves training two twin networks with shared weights: a high-resolution network and a low-resolution network. The high-resolution network serves as a teacher, guiding the learning process of the low-resolution network through feature map staggered alignment. The knowledge from the high-resolution network enhances the performance of the low-resolution network during low-resolution inference. Additionally, we employ a normalized skeleton loss to capture the loss of bone-related structure during training. Through extensive experiments on the MS-COCO and CrowdPose datasets, we demonstrate the superiority of our proposed method over state-of-the-art, lightweight multiperson pose estimation techniques, achieving much better performance with lower computational costs. Furthermore, our method achieves comparable performance to recent advanced regression-based pose estimation methods but with only 1/4 of the computational cost.**

***Index Terms*—2D pose estimation, lightweight neural networks**

I. Introduction

POSE estimation, a fundamental task in computer vision, accurately predicts the coordinates of the key points for individuals in an image. This critical task has gained significant prominence in recent years due to its wide range of applications, such as action recognition and motion analysis. As a result, pose estimation has become an indispensable component in various computer vision systems and contributes to advancements in diverse fields.

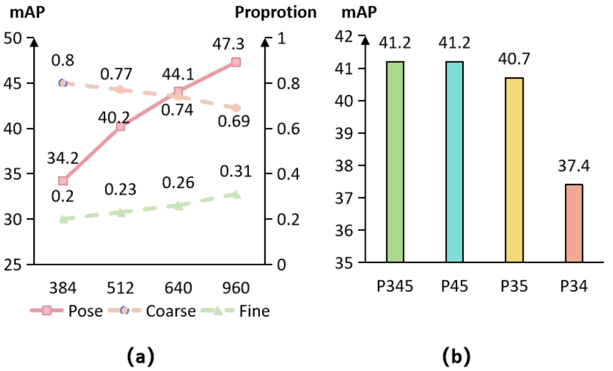


Fig. 1. We adopt a multiscale regression network structure for training and testing, where each scale in the PAN structure is treated as a level. (a) Model training and testing are performed with different resolutions and the output proportions from different levels after NMS. (b) Feature map-level contribution analysis is conducted. P_{ij} indicates that the model uses both the i^{th} level and j^{th} level for prediction.

Multiperson pose estimation can be divided into two types: top-down methods and single-stage methods. Top-down methods [41, 39] rely on a person detector to detect people in an image. The detected instances are then cropped and passed onto a single-person pose estimation network. Despite their high performance, the required human detectors are computationally expensive for lightweight multiperson pose estimation tasks. Single-stage methods can directly infer the coordinates of the keypoints for each person in an image without the need for additional models. Most of the current single-stage, multiperson pose estimation methods are heatmap-based approaches [8, 4], which involve learning the heatmaps of keypoints and subsequently determining their coordinates by finding the location of the extreme value points in the heatmaps. This approach is adopted by popular lightweight multiperson pose estimation methods [29, 38]. However, the limited computational resources available for lightweight models prevent them from producing large heatmaps. Therefore, the overall estimation accuracy of this type of algorithm is often unsatisfactory.

In contrast, regression-based multiperson pose estimation methods [24, 48] offer a promising approach for designing lightweight models. These approaches completely discard the

This work was supported by the Jiaxing Key Laboratory of Visual Big Data and Artificial Intelligence, Zhejiang Province, China Key R&D Program of Zhejiang Province (2021C01016)

(Corresponding author: Jungong Han, Tao Xu)

Zhenkun Fan and Zhuoxu Huang are with the Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, U.K. Zhenkun Fan is also with Shanghai Investigation, Design & Research Institute Co., Ltd. and AMATUS TECHNOLOGY Ltd. Zhuoxu Huang is also with Zhejiang Future Technology Institute and Taizhou Baite Technology Ltd. (e-mail: zhf1@aber.ac.uk; zhh6@aber.ac.uk).

Zhixiang Chen and Jungong Han are with the Department of Computer Science, The University of Sheffield, Western Bank Sheffield, S10 2TN U.K. (e-mail: zhixiang.chen@sheffield.ac.uk; jungonghan77@gmail.com).

Tao Xu is with the Shanghai Investigation, Design & Research Institute Co., Ltd. (e-mail: xt123@outlook.com).

Josef Kittler is with the Department of Electrical Engineering, Surrey University, Surrey GU2 7XH, U.K. (e-mail: j.kittler@surrey.ac.uk).

heatmap but directly regress the keypoint coordinates. With the widespread adoption of mature object detection networks[24, 25, 5, 22], the regression-based method from the YOLO series [37] has become increasingly prevalent for pose estimation. Leveraging these well-established object detection networks has led to promising results in the field of pose estimation.

Although regression-based methods show great potential for lightweight models due to their ability to bypass the need for generating large heatmaps, they have not yet been deployed to build lightweight pose estimation models primarily due to the 'resolution dilemma'. More specifically, training an accurate regression-based pose estimation model requires high-resolution images as input. It is evident from the latest regression-based works [25, 24] in which the use of high resolutions, e.g., 1280 ×1280 and 960 ×960, at both the training and testing stages is necessary. However, when we switch to low-resolution images, the performance of such models dramatically drops. On the other hand, in the context of the lightweight model, where computational resources are limited, the use of high-resolution images, especially at inference, does not seem practical. This finding raises a fundamental research question: How can we avoid performance decline when using low-resolution images within the constraints of a lightweight model?

To investigate the underlying reasons behind the performance degradation of the pose estimation model as the resolution decreases, we further explore into this issue. We believe that gaining insights into these factors will pave the way for effective solutions. With this aim, we conduct an extensive experiment employing a multiscale regression network architecture, wherein the PAN [21] structure encompasses three distinct feature map levels, each corresponding to a different output scale. Additionally, we train and evaluate this model using varying resolutions. During the evaluation, we examine the proportion of nonmaximum suppression (NMS) results that fall into each feature map level. Specifically, we designate the coarsest feature map level (P_3) as the coarse level, while the remaining levels (P_4 and P_5) are considered fine levels. As illustrated in Figure 1(a), we observe a distinct trend of the model favoring the use of the coarse level as the input resolution decreases. This phenomenon can be attributed to the varying receptive fields of the various feature map levels. When confronted with high-resolution inputs, the model demonstrates a preference for the finer level, which possesses a larger receptive field. However, in the case of low-resolution inputs, the model tends to rely more on the coarser level, which is characterized by a smaller receptive field.

In our investigation of feature map level performance, as depicted in Figure 1(b), we conduct a systematic analysis by selectively disregarding output feature maps from different levels during inference to assess their respective impacts on model performance. Through our experiments, we observe notable performance disparities among the feature map levels when employing regression networks for pose estimation tasks. Specifically, it became apparent that the performance contribution of the coarse level (P_3) consistently lagged behind that of the fine levels (P_4 and P_5). This discrepancy in performance is further amplified by the model's inherent tendency to favor higher-performing fine levels during high-resolution training and lower-performing coarse levels during low-resolution training.

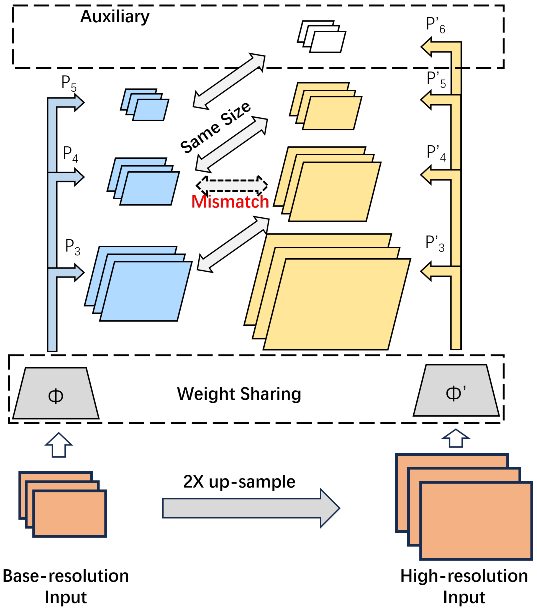


Fig. 2. Overview of our staggered alignment self-distillation method, which is designed to improve the performance of regression networks under low-resolution inputs. Our approach leverages self-distillation, wherein a high-resolution trained model, denoted as ϕ', serves as a teacher to guide a shared-weight model, ϕ, that is trained at a lower resolution. To ensure supervision between networks of different resolutions, we employ our staggered alignment strategy.

To counteract the performance degradation stemming from low-resolution inputs and to rectify observed performance discrepancies arising from resolutions, the concept of self-knowledge distillation has emerged as an auspicious solution. This technique serves as a conduit for harnessing the knowledge acquired from training the network with high-resolution inputs to inform the training of the same network using low-resolution inputs, thereby enhancing its performance in low-resolution inference scenarios. Furthermore, the self-distillation approach shares a single network as both the teacher and the student, which eliminates the need for an extra teacher network and multiple training steps. Additionally, we aim to exploit the superior performance exhibited by fine-level feature maps to guide the learning of coarse-level feature maps and to eventually alleviate performance disparities between them.

Nevertheless, self-distillation encounters two primary challenges. First, when introducing distillation learning from inputs of disparate resolutions, the disparity in feature map sizes resulting from high-resolution and low-resolution inputs complicates the direct distillation learning process. Second, because the network assigns distinct scale targets to different levels, the fine and coarse levels pursue separate learning objectives, rendering it impractical for the fine level to directly teach the coarse level learning process.

To address these two challenges, we propose our staggered alignment self-distillation (SASD) training method. As shown in 2, we adopt two identical networks with shared weights. These two networks use the same images but different resolutions as input. The high-resolution network uses an input that is upsampled by 2 times the base-resolution input. Consequently, each feature map in the high-resolution network is twice the size of the feature map in the base-resolution network at the same level, so the feature map size of the i^{th} level in the base-resolution network aligns with the more refined $(i+1)^{th}$ level in the high-resolution network. Existing multiscale learning methods dictate distinct learning objectives for different levels of the PAN network. In the training process, if the same target assignment approach is applied between networks ϕ and ϕ', the learning targets for P_i and P'_i would be identical, resulting in disparate objectives for P_i and P'_{i+1}. Considering that our previously introduced staggered alignment matching method has already matched P'_{i+1} to P_i for distillation, a discrepancy in learning objectives would degrade or even reverse the effectiveness of distillation. To address this issue, we adopted distinct label assignment strategies for networks ϕ and ϕ' during the implementation of staggered alignment self-distillation, and we assigned identical learning targets to P_i and P'_{i+1}. However, this approach introduces a complication in which the last level of the base resolution loses its matching counterpart. Thus, we introduce an auxiliary training level when using high resolution and align it with the last level of base resolution to allocate identical learning objectives.

Notably, this auxiliary level independently operates within the PAN structure and is exclusively utilized during high-resolution training. As a result, during the inference stage, the auxiliary level can be safely discarded without adversely affecting the overall performance. We refer to this approach as pruned inference, which allows for more efficient use of resources and improved performance.

In addition to our SASD method, we also tend to improve the regression-based pose estimation in the regression loss function. Traditionally, keypoint coordinate regression loss functions only calculate losses for individual keypoint positions, commonly utilizing $L1$ or $L2$ loss functions. Recent works such as YOLO-Pose [24] and YOLOv8 [36] have introduced the object keypoint similarity (OKS) loss function, which builds upon $L2$ loss by assigning different weights to each keypoint based on their importance. This weighted approach significantly enhances the training of regression methods. However, existing regression loss functions fundamentally focus on the distance between the predicted values and the ground truth on a point-by-point basis. For the human body, each keypoint is not isolated; therefore, capturing this correlation in the loss function could positively impact training. To address this issue, we propose a novel skeleton loss function that considers the difference in skeleton lengths between the predicted values and the GT as the loss. This skeleton loss function is designed to model the interdependency between two keypoints. Each adjacent pair of keypoints in the human body structure is considered to form a skeleton, and by using this loss function, the model learns the interdependency information between the keypoints connected by these skeletons. However, our current skeleton loss function is solely used to determine the length of each skeleton between the model outputs and labels without directly supervising the relative positions of each keypoint. This supervision of relative keypoint positions is enforced through the OKS (object keypoint similarity) loss function. During network training, both the OKS loss function and the skeleton loss function are jointly utilized to ensure simultaneous learning of both keypoint position information and the relationships among keypoints through the skeleton modeling process. In such scenarios, images contain targets of varying sizes, which leads to inherent imbalances in the differences. Larger objects naturally yield larger differences, which causes an imbalance in the loss between different-sized objects. To mitigate this imbalance, we normalize the loss function by the scale of the object, providing a solution to the size-related imbalance phenomenon, which we refer to as the normalized skeleton loss function.

Overall, we make the following contributions:

- First, we conduct an in-depth analysis of the resolution dilemma faced by regression-based multiperson pose estimation methods. To address this issue, we propose a novel training method, staggered alignment self-distillation training, and a dedicated network structure for distillation. This is the first attempt to employ self-distillation to address the resolution dilemma in pose estimation. Our approach effectively improves the performance of the regression model during low-resolution inference while maintaining computational efficiency.
- Second, diverging from conventional keypoint coordinate regression loss functions, which solely compute losses for individual keypoint positions, we introduce a normalized skeleton loss function. This innovation allows our model to grasp the interconnections among different key points of the human body, enhancing the precision of pose estimation.
- Last, we present experimental results on the MS-COCO [20] and CrowdPose [16] datasets, which demonstrate the effectiveness of our proposed method. Notably, our method outperforms current popular lightweight pose estimation methods. We achieve comparable performance to recent regression-based methods with only 1/4 of the computational cost.

II. Related Works

A. Top-Down Multiperson Pose Estimation

Top-down methods [17, 41, 6, 28, 34, 11, 15] are also described as two-stage approaches, in which an object detection network is employed to detect people in the image, and then each person is cropped and sent to a single-person pose estimation network to obtain their keypoints. SimpleBaseline [41] proposes a simple yet effective baseline network for top-down pose estimation by learning heatmaps of the body keypoints in each person and using deconvolution to increase the size of the output heatmap. Alternatively, Mask R-CNN [11] proposes a novel approach that solves the pose estimation problem through segmentation by predicting body keypoints using masks. In recent studies, the PCT [9] method has

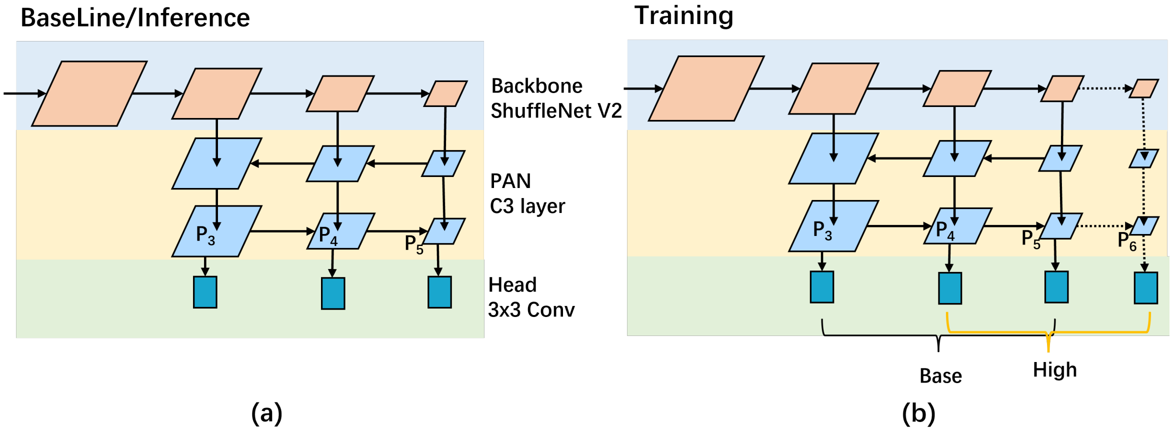


Fig. 3. Network architectures. (a) The baseline network architecture. The network used for training is denoted as (b). We introduce an additional finest P_6 as an auxiliary training level, which is subsequently discarded after training. As a result, the network structure during inference remains unchanged from the baseline, which ensures that no additional computational cost is introduced.

pioneered a new structured representation, which learns the presentation between body poses and tokens. Human pose estimation is transformed into a classification task that predicts token categories. Additionally, the BUCTD [47] introduces a hybrid top-down approach, which is distinct from previous top-down methods in which the first stage utilizes a bounding box detector. In contrast, the first stage of the BUCTD [47] method applies a single-stage pose estimation model to provide prior information for the second stage, thereby enhancing the model's performance in crowded scenarios. ICON [42] introduces an interimage contrastive consistency method that leverages contrastive learning to enhance the consistency of keypoint features between two images in the pose estimation task. This approach can be applied to improve the performance of various top-down methods. While top-down methods generally achieve superior performance, they require two different models and cannot be trained end-to-end. Additionally, the inference time of such methods increases linearly with the number of human bodies in the image, making them unsuitable for lightweight multiperson pose estimation.

B. Single-stage multiperson Pose Estimation

Single-stage multiperson pose estimation approaches can be divided into two categories: bottom-up methods and regression-based methods. The mainstream approach is bottom-up, which was originally proposed by Pishchulin in Deepcut [31] and was greatly improved in OpenPose [4] by Cao. These methods detect all human keypoints in an image at once and cluster them into persons. Most bottom-up methods [4, 27, 7, 14, 3] are based on heatmaps, where OpenPose generates heatmaps of the targets of each keypoint. OpenPole[4] uses Part Affinity Fields for keypoint clustering. Associative embedding [27] proposes additional vector embedding as a grouping method for keypoints, while HigherHRNet [7] improves model performance by producing high-quality and high-resolution heatmaps.

Recently, regression-based methods have gained popularity, with CenterNet [48] proposing a hybrid approach that regresses the center point of the human body using a heatmap and then regresses the coordinates of keypoints by the center point of the human body in feature maps. Due to the similarities between object detection and pose estimation tasks and significant advancements in object detection networks, recent works [24, 25] have attempted to use YOLO series object detection networks for pose estimation. These approaches discard the heatmap from the bottom-up method and use regression models to directly regress the keypoint coordinates, achieving particularly good results. While YOLO-like methods offer lightweight models, maintaining high performance necessitates higher input resolutions, thereby substantially increasing computational requirements. In recent research within the field of single-stage regression-based multiperson pose estimation, methods based on transformers have also begun to be widely applied. PETR [33] was the first to introduce the transformer into the domain of single-stage, regression-based multiperson pose estimation. This method incorporates the concept of object queries from the object detection domain, enabling NMS (nonmaximum suppression)-free methods in multiperson pose estimation. ED-Pose [45] employs a query selection approach. Initially, numerous coarse humans with a substantial number of object queries are generated, and then those with high confidence are filtered out for the next round of iterative local keypoint refinement, significantly enhancing the network's performance. Although transformer-based methods do not rely on high-resolution input, these transformer-based multiperson pose estimation methods face notable challenges, primarily due to the substantial computational complexity of the transformer structure, which results in slow processing speeds, making it challenging to achieve lightweight modifications.

Our proposed method can effectively address the main shortcomings of regression-based multiperson pose estimation models. Our staggered alignment self-distillation (SASD) method enhances the performance of lightweight models at lower resolutions through interlevel displacement self-distillation between different resolution inputs, thus enabling superior performance with reduced computational demands.

C. Lightweight Multiperson Pose Estimation

In the context of lightweight multiperson pose estimation, most current methods are lightweight versions of large bottom-up models. For instance, Lightweight OpenPose [29] proposes a lightweight backbone network and a lightweight head architecture to reduce the computational costs of OpenPose [4]. Moreover, EfficientHRNet [26] extends the EfficientNet [35] approach to HRNet [13] by combining reduced input resolution, a high-resolution network, and a heatmap prediction network. To further reduce the complexity of HRNet [13], LitePose [38] employed a network architecture search method to optimize the HRNet model. While regression-based methods have shown promising results on large models [24, 25], their applicability in lightweight multiperson pose estimation is constrained by the high input resolution necessary to achieve high performance. Thus, this paper addresses the challenges of applying regression-based methods to the lightweight pose estimation field.

D. Self-distillation

Knowledge distillation, which was initially proposed by Hinton [12], refers to the technique of transferring knowledge from a better-performing teacher model to a low-performance student model. Knowledge distillation can be categorized into three distinct types: offline distillation [12, 32], online distillation [1, 43], and self-distillation [46, 2, 44]. In offline distillation, the student model undergoes distillation learning by leveraging a pretrained teacher model. Conversely, online distillation entails a training paradigm in which both the teacher model and student model actively participate, facilitating concurrent parameter updates. Notably, self-distillation is a specific form of knowledge distillation in which the teacher and student models comprise the same network, eliminating the need for a separate, larger teacher model. Zhang [46] first proposed the self-distillation method, which enables the distillation of knowledge from the deeper parts of the network to the shallow parts of the network. Snapshot distillation [44], a variation of self-distillation, involves transferring knowledge from earlier epochs (teachers) to later epochs (students) within the same network, enabling a supervised training process. OKDHP [19] employs a self-distillation approach to utilize the knowledge aggregated from multibranch learning to guide the learning of each individual branch. The authors propose a feature aggregation unit (FAU) to aggregate the heatmaps generated by each branch of the multibranch network and use the aggregated heatmap to supervise each single branch. DSKD [40] employs a densely guided self-knowledge distillation framework to solve the error avalanche problem in multiteacher distillation, which enhances the quality of heatmaps in the scenario of multiteacher distillation.

Previous self-distillation approaches primarily focus on self-supervised learning across different stages or various outputs of the model, typically utilizing the same input resolution during training. In contrast, our self-distillation method addresses the performance gap during training with varying resolutions. By aligning feature maps generated at different levels from high-resolution inputs with those from low-resolution inputs, our method effectively addresses the challenge of guiding learning from high-resolution inputs to low-resolution inputs, consequently enhancing model performance at low resolutions. Remarkably, self-distillation methods have yet to be applied to enhance model performance in low-resolution inference. Thus, in this study, we employ, for the first time, self-distillation as a means to address this challenge.

III. Methodology

In this section, we present our lightweight regression-based pose estimation network. First, we introduce an additional finest auxiliary training level that operates independent of other levels, ensuring its autonomy and effectiveness in our self-distillation training.

Second, to address the challenge of low-resolution inference, we present our new SASD method. This approach involves utilizing two networks with varying resolution inputs that share weights, with the high-resolution network serving as a guide for the learning process of the base-resolution network. Subsequently, we apply the pruned inference, which allows us to discard the auxiliary level during low-resolution inference. By doing so, we effectively maintain the model the same as the baseline model, eliminating any redundant computational burden associated with the auxiliary training level.

Last, we introduce the comprehensive loss function in our training approach, which incorporates our novel normalized skeleton loss. This loss function effectively captures the correlation between human keypoints, leading to improved performance of our network.

A. Network Architecture

Our baseline lightweight pose estimation network architecture is built upon the YOLO-style architecture, as shown in Figure 3 (a). ShuffleNetV2 [23] is employed as the backbone network, followed by three PAN [21] layers with 3 different scales for effective multiscale feature fusion. These features are then passed through three regression heads to estimate the coordinates of the bounding boxes and keypoints, as well as the confidence scores for each keypoint. We use 1×1 convolution and C3 layers from YOLOv5 in the PAN layers and 3×3 convolution in the heads.

In our PAN structure, we denote the three different level scales as P_3, P_4, and P_5. We limit the number of channels and layers in our model to ensure its lightweight nature, resulting in a total of only 1.6 M parameters. This design makes our model suitable for real-time applications on resource-constrained mobile devices or embedded systems. Furthermore, this parameter count allows for a fair comparison with recent lightweight pose estimation methods [38].

During the training process, we introduce an auxiliary training level, which consists of a ShuffleBlock and two C3 layers, which are denoted as P_6, as shown in Figure 3 (b), in our baseline model for SASD training. Unlike the other feature map levels, P_6 remains decoupled from the other levels within the feature pyramid structure. For the distillation training,

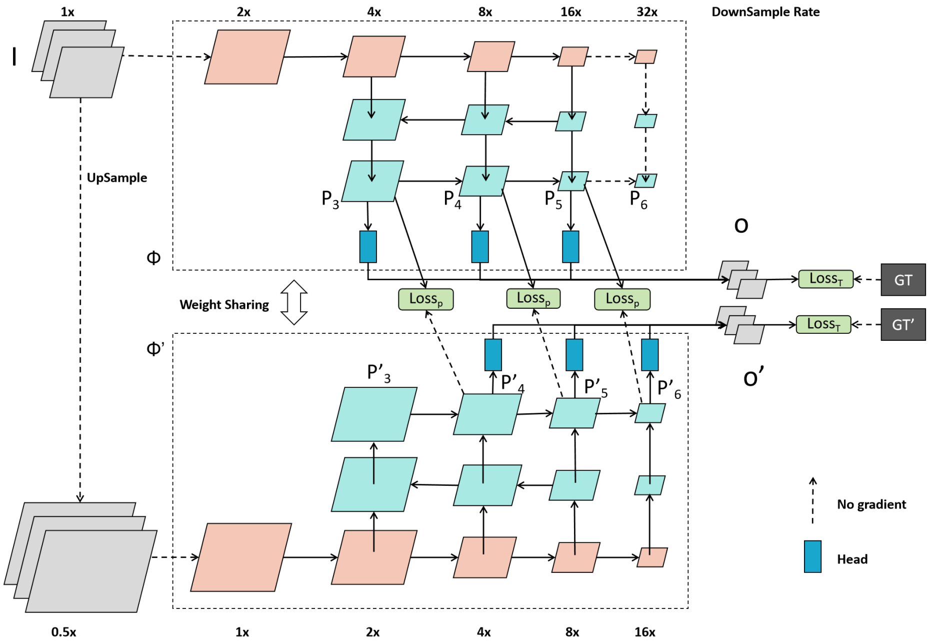


Fig. 4. Staggered alignment self-distillation. $P_3 - P_6$ represents three different feature map levels with different scales. We utilize feature maps of P' obtained from high-resolution inputs to supervise the feature maps P obtained from low-resolution inputs. $Loss_P$, which represents the L2 loss, is computed between P and P'. $Loss_T$ represents the combination of OKS loss and the normalized skeleton loss function.

we ensure that P_3, P_4, P_5, and P_6 have the same channel dimensions.

After training, P_6 is discarded to maintain the original network structure during inference, thereby avoiding any additional computational cost. We name this approach pruned inference.

B. Staggered Alignment Self-distillation

Here, we explain we adopt self-distillation, namely, staggered alignment self-distillation, to improve the performance of our model during low-resolution inference. We present a novel training method called staggered alignment self-distillation. Basically, this method utilizes two twin networks with shared weights but operates at different resolution inputs. The high-resolution network provides supervision to guide the learning process of the low-resolution network.

As illustrated in Figure 4, network ϕ operates at the base resolution and shares its weights with network ϕ' operating at the high resolution. During the learning process, we adopt the label assignment strategy from YOLOv5 to assign targets of varying scales to the P_3, P_4, and P_5 levels of network ϕ. Simultaneously, we apply the same label assignment strategy to the P'_4, P'_5, and P'_6 levels of network ϕ', ensuring consistent target assignments between the P_i level and the P'_{i+1} level. Due to the utilization of weight sharing in our approach, both P_4 and P'_4, as well as P_5 and P'_5, receive supervision from two distinct target scales. This approach augments the training samples for P_4 and P_5. The auxiliary training level P_6 does not engage in the training of base-resolution inputs, but P'_6, which shares weights with it, participates in the training process for high-resolution inputs.

The input I' of network ϕ' is derived from the original input I through bilinear upsampling, which results in the feature maps of network ϕ' always being twice the size of the corresponding feature maps in ϕ. Additionally, owing to the identical channel configuration across all these levels, the feature map P'_{i+1} of network ϕ' and the feature map Pi of network ϕ possess identical spatial dimensions.

Our previous analysis revealed that networks trained at higher resolutions outperform those trained at lower resolutions. Additionally, finer levels have already demonstrated superior performance compared to coarser levels. To address this issue, we leverage feature map supervision, which is commonly employed in knowledge distillation. In particular, we employ the feature map P'_{i+1} of the high-resolution network ϕ' to supervise the feature map P_i of the base resolution network ϕ. The loss function is expressed as follows:

$$L_P(P, P') = \sum_{i=3}^{5} \|P'_{i+1} - P_i\|_2. \tag{1}$$

In this equation, we compute the $L2$ distance between P'_{i+1} and P_i.

In addition to learning between feature maps, we also apply the OKS loss function L_{OKS} and normalized skeleton loss L_{SK} function for keypoint regression at both the base resolution and higher resolution. The object predictions of the base-resolution network ϕ are learned through its P_3, P_4, and P_5 levels, while the high-resolution network learns predictions through its P'_4, P'_5, and P'_6 levels. This finding can be represented by the following equation.

$$O = H(P_3, P_4, P_5), \tag{2}$$

$$O' = H'(P'_4, P'_5, P'_6), \tag{3}$$

TABLE I
RESULTS ON THE MS-COCO [20] VAL/TEST-DEV SET AND CROWPOSE [16] TEST SET. COMPARED TO OTHER LIGHTWEIGHT POSE ESTIMATION METHODS, WE HAVE ADVANTAGES IN TERMS OF THE NUMBER OF PARAMETERS, COMPUTATIONAL COST (MACS), AND LATENCY. ALL LATENCIES ARE TESTED ON THE QUALCOMM SNAPDRAGON 855. FOR FAIR COMPARISONS, WE KEEP OUR MACS AT THE SAME LEVEL AS LITEPOSE BY ADJUSTING INFERENCE RESOLUTION.

Method	Params ↓	MACs ↓	Latency(ms) ↓	AP$_{val}$ ↑	AP$_{test-dev}$ ↑
MS-COCO					
PersonLab [30]	68.7M	206.9G	-	-	56.6
HigherHRNet-W48 [7]	63.8M	155.1G	1532	69.9	68.4
SRPose [10]	23.5M	30.86G	-	48.4	-
Lightweight OpenPose [29]	4.1M	9.0G	97	42.8	-
EfficientHRNet-H$_{-2}$[26]	8.3M	7.9G	182	52.9	52.8
LitePose-S [38]	2.7M	5.0G	76	56.8	56.7
YOLOv8-N [36]	3.3M	4.35G	66	50.5	-
ours-E	2.4M	5.0G	78	57.2	56.9
EfficientHRNet-H$_{-4}$[29]	2.8M	2.2G	78	35.7	35.5
LitePose-XS [38]	1.7M	1.2G	27	40.6	37.8
MFite-HRNet [18]	1.8M	2.43G	-	41.4	-
ours-XS	1.6M	1.2G	24	44.1	43.7
CrowdPose					
Scaled-HigherHRNet-W16 [7]	7.2M	12.5G	170	-	50.4
EfficientHRNet-H$_{-3}$[29]	5.3M	4.3G	132	-	46.1
LitePose-XS [38]	1.7M	1.2G	27	-	49.5
ours-XS	1.6M	1.2G	24	-	50.4

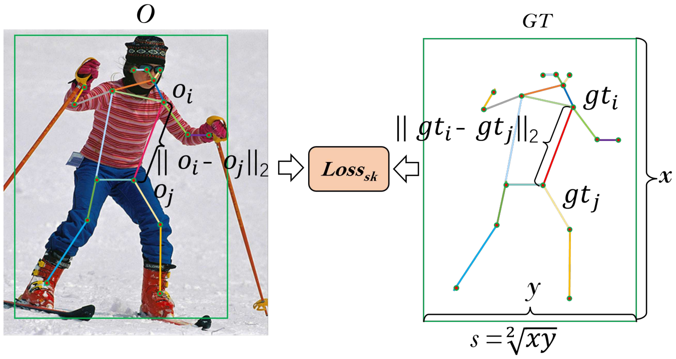


Fig. 5. Calculation of the normalized skeleton loss function, where o denotes the network's output for keypoint coordinates, gt represents the ground truth labels for keypoint coordinates, and s signifies the scaling factor utilized for normalization, as computed by the x and y in the bounding box of the human body.

where H and H' represent the head structures in networks ϕ and ϕ', O and O' correspondingly represent the output results of ϕ and ϕ', utilized for the computation of the loss function. Notably, we obtain high-resolution images using bilinear interpolation from the original input without acquiring high-resolution images from the dataset during training. This step is important because high-resolution images can be difficult or expensive to obtain, and our approach avoids this need. Moreover, during our interfeature map supervision training process, the performance of our network's feature maps is not good enough at the beginning of the training. To avoid the adverse impact of low-quality feature maps on network performance during training from scratch, we need to perform a warm-up phase for a certain number of epochs before using interfeature map supervision. Next, we can continue with SASD on our model.

C. Loss Function

In the training phase, we employed the OKS loss function [24] for keypoint regression, which is defined in the following equation.

$$L_{OKS} = 1 - \frac{\sum_{i=0}^{N_{kpts}-1} exp(\frac{(\|o_i - gt_i\|_2)^2}{2s^2k_n^2})\delta(v_n > 0)}{\sum_{i=0}^{N_{kpts-1}} \delta(v_n > 0)}, \quad (4)$$

k_n represents the weight of keypoints; s represents the scale of the person; N_{kpts} represents the number of keypoints in the human body; o_i and gt_i denote the output result and corresponding label for the coordinate of the i^{th} keypoint, respectively; and $\delta(v_n > 0)$ is the visibility flag for each keypoint.

In addition, we propose a novel normalized skeleton loss function that calculates the normalized skeleton distance between the predicted result and the ground truth (i.e., L2 distances between keypoints) to learn the interdependence relation of the keypoints. The normalized skeleton loss function is represented by the following equation:

$$L_{SK} = \frac{\sum_{i,j \in set_{sk}} \|\|o_i - o_j\|_2 - \|gt_i - gt_j\|_2\|_2}{s}, \quad (5)$$

where set_{sk} denotes the set of all adjacent keypoints of the human body. To maintain similar loss distributions for objects of different sizes, we divided the results by the scale of the object s, as depicted in Figure 5.

In addition, we use L_P to calculate the distance between the feature maps for different input sizes, as shown in equation 1. The loss function for training can be represented by the following equation.

$$\begin{aligned} L = \alpha(L_{OKS}(O', GT') + L_{OKS}(O, GT)) \\ +\beta(L_{SK}(O', GT') + L_{SK}(O, GT)) \\ +\gamma L_P(P, P'), \end{aligned} \quad (6)$$

O and GT denote the output prediction and ground truth, respectively, for images of the base resolution. Conversely,

TABLE II

Comparison with other regression-based multi-person pose estimation methods on MS-COCO [20] val/test-dev set. Compared to other regression-based methods, we have great advantages in computational cost (MACs) and latency. All latencies are measured using an NVIDIA RTX3090Ti GPU. *The multi-scale inference strategy is employed in the testing of KAPAO [25].

Method	Input size	MACs ↓	Latency(ms) ↓	$\text{AP}_{val} \uparrow$	$\text{AP}_{test-dev} \uparrow$
YOLOPose-S [24]	640	10.2G	5.5	57.0	-
YOLOPose-S[24]	960	22.8G	11.9	63.8	62.9
Kapao-S [25]	1280	40.5G	21.7	63.0	63.8*
YOLOv8-S [36]	640	15.1G	7.8	60.2	-
ours-S	640	10.1G	5.3	63.5	62.2
YOLOPose-M [24]	960	66.3G	23.0	67.4	66.6
Kapao-M [25]	1280	117G	43.5	68.5	68.8*
YOLOv8-M [36]	640	40.5G	20.0	65.0	-
ours-M	640	27.7G	10.6	68.6	67.0
YOLOPose-L [24]	960	145.6G	34.0	69.4	68.5
Kapao-L [25]	1280	258.7G	63.3	70.6	70.3*
YOLOv8-L [36]	640	84.3G	18.7	67.6	-
ED-Pose [45]	1066	144.6G	-	69.9	-
ours-L	640	61.7G	16.4	70.1	68.8

O' and GT' represent the prediction value and ground truth value, respectively, for upsampled images. The ground truth GT' is obtained by scaling up GT using the upsampling rate as the magnification factor. The variables α, β, and γ are the balance parameters of the equation. In our experiments, we set $\alpha = 0.1$, $\beta = 0.1$, and $\gamma = 0.02$.

IV. Experiments

We evaluate our model on the MSCOCO [20] and CrowdPose [16] datasets.

MS COCO [20]. The MS COCO dataset comprises more than 200,000 images of the human body, each of which contains 17 keypoints. The dataset is segregated into three subsets: a training set, a validation set, and a test set, consisting of 57K, 5K, and 20K images, respectively. We trained our experiments on the MS COCO dataset exclusively using the training set. Our paper reports the performances of the models on the validation set and test set.

CrowdPose [16]. The CrowdPose dataset encompasses more than 20,000 images of the human body, annotated with 14 keypoints. We follow the preceding methodologies [7, 38] and train our models on both the training set and validation set to showcase the results obtained on the test set.

Metrics. Our test metrics are based on the object keypoint similarity (OKS), and the results we report in the paper are the mean accuracy (mAP) of the OKS.

A. Experiment Settings

During the training stage, we employ augmentation strategies similar to those used in [24]. Specifically, we utilize random translation in the range of [-10, 10], random flipping with a probability of 0.5, mosaic augmentation with a probability of 1, and an array of color augmentations. We use the SGD optimizer with a cosine scheduler. Our base learning rate is set to 1e-2. In the training process, first, we conduct a warm-up process by training the model for 100 epochs at the base resolution. Second, we train each model for 300 epochs using our SASD approach. For testing, first, we resize the larger side of the input images to the desired size while maintaining the aspect ratio. Second, we pad the lower side of the image to generate a square image, ensuring that all the input images are of the same size; we also use the flip test of LitePose [38].

B. Main Results

Table I presents the experimental results of our model on the MS-COCO and CrowdPose datasets. We report our XS version as shown in Figure 3. Additionally, to ensure a fair comparison with other methods, we increase the number of feature map levels to 4 to enlarge our model to our E version.

Our experimental results show a significant reduction in computational cost and the number of parameters while achieving much better performance than other lightweight pose estimation methods, such as Lightweight OpenPose [29], EfficientHRNet [26] and SRPose[10]. Notably, our proposed model demonstrates a substantial performance advantage over the LitePose [38] approach, achieving 5.3 mAP higher results on the Ms-COCO test set. Remarkably, our model maintains fewer parameters, has lower computational costs, and enables faster inference on mobile devices. Furthermore, our method also outperforms MFite-HRNet[18] by 2.7 mAP on the MS-COCO validation dataset with fewer parameters and MACs.

Additionally, we compare our method with recent regression-based multiperson pose estimation methods, including YOLOPose[24], KAPAO[25] and YOLOv8 [36]. Similar to YOLOPose[24], KAPAO[25] and YOLOv8[36], we design three network versions based on YOLOv5S, YOLOv5M, and YOLOv5L, respectively, and train them using our staggered alignment self-distillation. As shown in Table II, when using the same resolution of 640, our approach outperforms the YOLO-pose[24] method by a significant margin, with a 5-point increase in the mAP. Our method also achieves better results than YOLOPose, which is trained at a higher resolution of 960×960 and performs comparably to KAPAO[25], which is trained at a higher resolution of 1280×1280. Despite our usage of a lower resolution of 640×640 for training, our method also achieves these results with only 1/4 of the computational cost of KAPAO[25]. In addition, our model outperforms the recent YOLOv8[36] method in terms of both

TABLE III
ABLATION EXPERIMENTS OF DIFFERENT TRAINING STRATEGIES

Method	Training Resolution	Inference Resolution	Result (mAP)
Baseline	512	512	40.3
Random Scale	[0.5,1.5]×512	512	39.5
High Resolution	1024	512	40.5
SASD	512,1024	512	44.1
SASD w/o weight sharing	512,1024	512	42.3
SASD w/o auxiliary training level	512,1024	512	43.2
Direct Distillation	512,1024	512	39.3

computational complexity and performance when operating at the same resolution of 640. Furthermore, we also conduct a comparison with the latest ED-Pose [45], which applies the transformer architecture for single-stage, multiperson pose estimation. The experimental results indicate that our model achieves better performance with a lower computational cost. This comparison validates the effectiveness of our approach.

C. Ablation Experiments

To evaluate the effectiveness of our proposed techniques, we conducted comprehensive ablation experiments on the MS-COCO dataset. These experiments involved training our network on the training set of the MS-COCO dataset and the performance of different methods evaluated on the validation set. The purpose of these ablation experiments is to isolate and analyze the impact of each individual component in our approach, enabling us to gain a deeper understanding of how they collectively contribute to improvement.

1) Ablation Experiments of Different Training Strategies: To evaluate the efficacy of different training strategies, we conducted a comparative experiment among various methods, as depicted in Table III. The baseline method employs a fixed resolution of 512×512 as the input during training. In contrast, the random scale method employs a random resolution ranging from 0.5 to 1.5 times the 512×512 resolution for training. The high-resolution method, on the other hand, adheres to a fixed resolution of 1024×1024 during the training phase. Our proposed staggered alignment self-distillation (SASD) method utilizes 512×512 resolution and incorporates an upscale to 1024×1024 resolution as input during training. To establish the necessity of weight sharing in our SASD approach, we conduct experiments wherein the SASD algorithm is employed without weight sharing. In this experimental setting, two networks trained with different resolutions are decoupled and no longer share weights. During the inference phase, the inference is performed solely using the network trained on low-resolution images. To verify the effectiveness of the auxiliary training level, we remove both the auxiliary training levels P_6 and P'_6 while keeping other settings similar to those of the SASD experiment in our SASD experiments w/o the auxiliary training level. In the direct distillation experiment, we did not employ the staggered alignment matching method for distillation. Instead, we perform distillation by using the network trained at high resolution to supervise the corresponding layer of a network trained at low resolution, i.e., using P'_i to supervise P_i. Due to mismatched feature map sizes, we downsample the feature maps generated by the high-resolution network by a factor of 2 to align with the low-resolution network. We present the results of all methods tested using a consistent 512×512 resolution on the MS-COCO validation set.

Moreover, during the inference stage, all methods employ identical model architectures, ensuring a fair comparison. The experimental outcomes illustrated in Table III indicate that adopting multiple random resolutions during training does not yield any performance improvements. Additionally, training with high-resolution data and subsequently testing on lower resolutions only results in marginal performance gains, as evidenced by the reported results. Remarkably, our proposed SASD method outperforms single-resolution and high-resolution methods, underscoring its effectiveness in enhancing performance.

Furthermore, our experiments highlighted the synergistic impact of weight sharing within networks, as evidenced by the noticeable decrease in performance of 1.8 mAP when weight sharing is omitted in the SASD experiment. This finding underscores the critical importance of weight sharing in our SASD approach. We attribute this phenomenon to the notion that weight sharing leads to different sample distributions for the network when addressing base-resolution and high-resolution inputs. Essentially, this approach provides more training samples for P_4 and P_5, effectively acting as data augmentation. Moreover, when examining the experimental results of the SASD without auxiliary training, we observe a performance decrease of 0.9 mAP compared to that in the experiments with auxiliary training. This observation confirms the significant positive influence that auxiliary training levels have on the model's performance and their effectiveness in enhancing the overall performance of the model. The experimental results of the direct distillation indicate that using downsampled high-resolution feature maps for direct distilling does not yield positive gains for the network; in contrast, it leads to a slight degradation in performance. This decline is attributed to the substantial loss of information incurred during the downsampling of high-resolution feature maps, which is detrimental to the distillation learning process. This finding indirectly validates the necessity of employing the SASD algorithm in our approach.

2) Ablation Experiments of Loss Functions: To further explore the impact of different loss functions on performance during staggered alignment self-distillation training, we conduct comprehensive experimentation by employing various combinations of loss functions. The corresponding model performance under different loss function combinations is

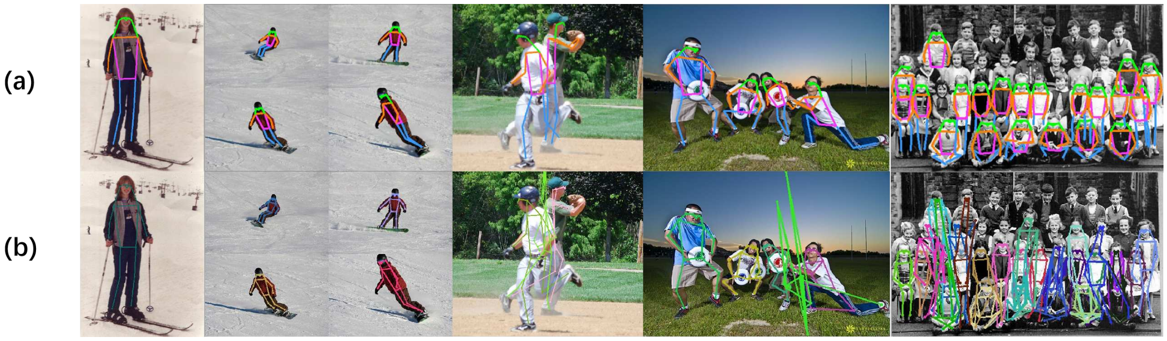


Fig. 6. Visualization of the prediction results: (a) shows the predictions of our model and (b) shows the predictions of litepose[38]

TABLE IV
ABLATION EXPERIMENTS OF DIFFERENT LOSS FUNCTIONS

Methods			Result (mAP)
OKS Loss	Skeleton Loss	Distillation Loss	-
✓			40.6
✓	✓		41.7
✓		✓	43.6
✓	✓	✓	44.1

carefully evaluated and analyzed.

During the experiments, we apply staggered alignment self-distillation training and pruned inference for all the models, where a base resolution of 512×512 is employed as the input, and two upsamplings are applied to obtain the high-resolution input. The performance evaluation is conducted on the MS COCO validation dataset at a resolution of 512×512.

The results presented in Table IV clearly indicate that our proposed skeleton loss significantly enhances the model's performance, exhibiting a substantial increase of 0.9 mAP compared to models solely relying on the OKS loss. Additionally, our distillation loss demonstrates its pivotal role, resulting in an impressive increase of 2.8 mAP when combined with the OKS loss. The synergistic combination of skeleton loss and distillation loss yields a remarkable improvement in the model's performance, highlighting their collective effectiveness in achieving superior results.

TABLE V
ABLATION EXPERIMENTS OF NORMALIZED SKELETON LOSS.

Method	Result(mAP)
Baseline	40.6
Normalized Skeleton Loss	41.7
Skeleton Loss w/o normalization	38.8

TABLE VI
ABLATION EXPERIMENTS ON DIFFERENT ALGORITHMIC COMPONENTS.

Methods		Result (mAP)	MACs	Latency(ms)
SASD	Pruned Inference	-	-	-
		40.3	1.2G	24
✓		43.8	1.9G	32
✓	✓	44.1	1.2G	24

3) Ablation Experiments of the Normalized Skeleton Loss: We conducted experiments as an ablation study to investigate the influence of normalization on the skeleton loss function, as shown in Table V. These experiments are conducted with the same experimental setup, using our skeleton loss function in the baseline network architecture. The key distinction lies in the absence of the normalization step in the final set of experiments. This deliberate variation allowed us to discern the impact of normalization on our proposed loss function.

By incorporating normalization, the skeleton loss exhibits a 1.1 mAP improvement compared to that of the baseline. Conversely, in the absence of normalization, the skeleton loss decreases by 1.8 mAP compared with the baseline. The experimental results indicate that the absence of normalization has a detrimental effect on model performance when the skeleton loss function is used. This finding confirms that skeleton loss contributes to network performance only when normalization is applied. Although calculating the loss between skeletons can capture the relationship between key points, addressing the imbalance in loss due to varying target sizes is a crucial aspect where skeletal loss can be effective in multiperson pose estimation.

4) Ablation Experiments on Different Algorithmic Components: In the final phase of our research, we conduct a series of ablation experiments to empirically evaluate the effectiveness of each component in our proposed methodology, as shown in Table VI. Specifically, we investigate the impact of the SASD method and the pruned inference technique by comparing combinations of these methods with the baseline network. Throughout our experiments, we maintain a fixed resolution of 512×512 for all models and report their inference results at this resolution.

The comprehensive results demonstrate that each proposed method significantly contributes to the overall performance improvement. Notably, when solely employing the SASD method without incorporating the pruned inference technique, the computational cost of the model increased to 1.9 G due to the introduction of an auxiliary training level, and the performance increased by 3.5 mAP. The effectiveness of the SASD method stems from our base resolution network, which can acquire knowledge from networks trained at higher resolutions. In addition, the coarse level of our network is also supervised by the fine level with higher performance, thereby achieving more precise recognition of small and medium-sized targets.

However, upon integrating the pruned inference technique, we successfully restored the model size to be on par with that of the baseline model. We introduced the auxiliary training level in the SASD network architecture for training. This level is exclusively utilized during training with high-resolution inputs. However, when operating at the base resolution, the corresponding level remains structurally present but inactive. During the inference phase, where the base resolution is used, the features generated by the auxiliary training branch not only fail to positively contribute but also may impede the network's operation. Given that the auxiliary training level is designed as an independent module during network architecture, other levels do not depend on its feature maps. Therefore, we employ the pruned inference method during inference to trim the auxiliary training level, enhancing network performance while reducing the computational cost.

D. Visualization of the predicted results

In this section, we present the visualization results of our method on the MS-COCO dataset and compare them with those of the lite-pose [38] method. As shown in Figure 6, our model performs consistently well in sparse multiperson scenarios, similar to lite pose, but we exhibit a significant advantage in dense crowd scenarios. This is attributed to the advantage that our regression model directly predicts the 17 keypoint coordinates for each person without the need for keypoint matching, as required by bottom-up methods. We eliminate the problem of incorrect keypoint matching from the source.

E. Inference time on the edge computing device

To assess the effectiveness of our model on edge computing devices, we conduct rigorous tests on an ARM A53 architecture-based CPU, which is a widely adopted low-power, low-computational-capacity CPU architecture. Our evaluation primarily focuses on the inference speed under varying core configurations.

For inference, we employ NCNN as our framework, ensuring compatibility with standard deep-learning libraries and facilitating seamless deployment across diverse platforms.

Our experiments shown in Table VII demonstrate remarkable performance even with a single core, achieving an impressive inference rate that exceeds 6 frames per second. Leveraging the power of four cores, our algorithm achieves a threefold speed improvement compared to the single-core setup, thereby enabling smooth and real-time pose estimation.

TABLE VII
EXPERIMENTS OF INFERENCE TIME ON THE EDGE COMPUTING DEVICE

Method	MACs	Inference Time(1xCore)	Inference Time(4xCore)
XS	1.2G	161.1ms	54.7ms
E	5G	308.6ms	97.3ms

V. CONCLUSION

In this paper, we analyzed the challenges faced by applying regression-based multiperson pose estimation methods to lightweight architecture, focusing on the resolution dilemma in regression-based networks when used for pose estimation tasks. To address these issues, we propose not only a staggered alignment self-distillation approach and its corresponding network architecture to improve the low-resolution inference performance of the network but also a normalized skeleton loss function to enhance the keypoint relationships and improve model performance. Our experiments on the MS-COCO and CrowdPose datasets demonstrated the merits of our proposed methods compared to recent lightweight pose estimation techniques, which achieved superior performance with fewer computational requirements. Additionally, our ablation experiments confirmed the efficacy of each of our proposed components, which validates their importance in achieving our overall results.

REFERENCES

[1] Rohan Anil et al. "Large scale distributed neural network training through online distillation". In: *arXiv preprint arXiv:1804.03235* (2018).

[2] Prashant Bhat, Elahe Arani, and Bahram Zonooz. "Distill on the go: online knowledge distillation in self-supervised learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2678–2687.

[3] Weixi Cai. "Improvement in Multi-Person 2D Pose Estimation: Applying Polar Representation in OpenPose". In: *2021 2nd International Conference on Computing and Data Science (CDS)*. IEEE. 2021, pp. 313–318.

[4] Zhe Cao et al. "Realtime multi-person 2d pose estimation using part affinity fields". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7291–7299.

[5] Changrui Chen, Jungong Han, and Kurt Debattista. "Virtual Category Learning: A Semi-Supervised Learning Method for Dense Prediction with Extremely Limited Labels". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[6] Yilun Chen et al. "Cascaded pyramid network for multi-person pose estimation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7103–7112.

[7] Bowen Cheng et al. "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5386–5395.

[8] Zigang Geng et al. "Bottom-up human pose estimation via disentangled keypoint regression". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14676–14686.

[9] Zigang Geng et al. “Human Pose as Compositional Tokens”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 660–671.

[10] Jie Tang Haonan Wang Jie Liu and Gangshan Wu. “Lightweight Super-Resolution Head for Human Pose Estimation”. In: *arXiv preprint arXiv:2307.16765* (2023).

[11] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015).

[13] Junjie Huang, Zheng Zhu, and Guan Huang. “Multi-stage HRNet: Multiple stage high-resolution network for human pose estimation”. In: *arXiv preprint arXiv:1910.05901* (2019).

[14] Lei Jin et al. “Grouping by center: Predicting centripetal offsets for the bottom-up human pose estimation”. In: *IEEE Transactions on Multimedia* (2022).

[15] Aouaidjia Kamel et al. “Hybrid refinement-correction heatmaps for human pose estimation”. In: *IEEE Transactions on Multimedia* 23 (2020), pp. 1330–1342.

[16] Jiefeng Li et al. “Crowdpose: Efficient crowded scenes pose estimation and a new benchmark”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 10863–10872.

[17] Qun Li et al. “HRNeXt: High-Resolution Context Network for Crowd Pose Estimation”. In: *IEEE Transactions on Multimedia* (2023).

[18] Shuo Li et al. “A lightweight pose estimation network with multi-scale receptive field”. In: *The Visual Computer* (2023), pp. 1–12.

[19] Zheng Li et al. “Online knowledge distillation for efficient pose estimation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 11740–11750.

[20] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.

[21] Shu Liu et al. “Path aggregation network for instance segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8759–8768.

[22] Yi Liu et al. “Part-object relational visual saliency”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3688–3704.

[23] Ningning Ma et al. “Shufflenet v2: Practical guidelines for efficient cnn architecture design”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 116–131.

[24] Debapriya Maji et al. “YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 2637–2646.

[25] William McNally et al. “Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation”. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*. Springer. 2022, pp. 37–54.

[26] Christopher Neff et al. “Efficienthrnet: Efficient scaling for lightweight high-resolution multi-person pose estimation”. In: *arXiv preprint arXiv:2007.08090* (2020).

[27] Alejandro Newell, Zhiao Huang, and Jia Deng. “Associative embedding: End-to-end learning for joint detection and grouping”. In: *Advances in neural information processing systems* 30 (2017).

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked hourglass networks for human pose estimation”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer. 2016, pp. 483–499.

[29] Daniil Osokin. “Real-time 2d multi-person pose estimation on cpu: Lightweight openpose”. In: *arXiv preprint arXiv:1811.12004* (2018).

[30] George Papandreou et al. “Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 269–286.

[31] Leonid Pishchulin et al. “Deepcut: Joint subset partition and labeling for multi person pose estimation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4929–4937.

[32] Adriana Romero et al. “Fitnets: Hints for thin deep nets”. In: *arXiv preprint arXiv:1412.6550* (2014).

[33] Dahu Shi et al. “End-to-end multi-person pose estimation with transformers”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11069–11078.

[34] Ke Sun et al. “Deep high-resolution representation learning for human pose estimation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5693–5703.

[35] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[36] Ultralytics. *YOLO V8*. https://github.com/ultralytics/ultralytics. 2023.

[37] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors”. In: *arXiv preprint arXiv:2207.02696* (2022).

[38] Yihan Wang et al. “Lite pose: Efficient architecture design for 2d human pose estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13126–13136.

[39] Shih-En Wei et al. “Convolutional pose machines”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016, pp. 4724–4732.

[40] Mingyue Wu et al. “Lightweight Human Pose Estimation Based on Densely Guided Self-Knowledge Distillation”. In: *International Conference on Artificial Neural Networks*. Springer. 2023, pp. 421–433.

[41] Bin Xiao, Haiping Wu, and Yichen Wei. “Simple baselines for human pose estimation and tracking”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 466–481.

[42] Xixia Xu et al. “Inter-image contrastive consistency for multi-person pose estimation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 3. 2023, pp. 3063–3071.

[43] Aijia Yang et al. “Context Matters: Distilling Knowledge Graph for Enhanced Object Detection”. In: *IEEE Transactions on Multimedia* (2023).

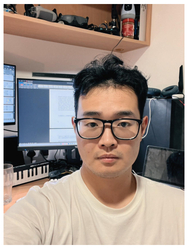
[44] Chenglin Yang et al. “Snapshot distillation: Teacher-student optimization in one generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2859–2868.

[45] Jie Yang et al. “Explicit Box Detection Unifies End-to-End Multi-Person Pose Estimation”. In: *International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=s4WVupnJjmX.

[46] Linfeng Zhang et al. “Be your own teacher: Improve the performance of convolutional neural networks via self distillation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3713–3722.

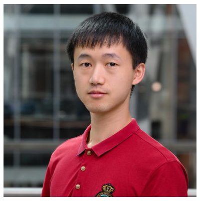
[47] Mu Zhou et al. “Rethinking Pose Estimation in Crowds: Overcoming the Detection Information Bottleneck and Ambiguity”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 14689–14699.

[48] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. “Objects as points”. In: *arXiv preprint arXiv:1904.07850* (2019).

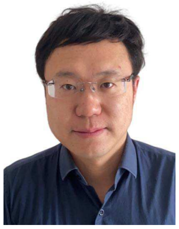


Zhenkun Fan , currently pursuing a Ph.D. degree in Computer Science at Aberystwyth University, UK. His research focuses on areas of computer vision such as Pose estimation and Sematic Segmentation, He received his master’s degree in computer science from Ocean University of China in 2022 and a B.E. degree in Computer Science from Shandong University of Science and technology in 2019.

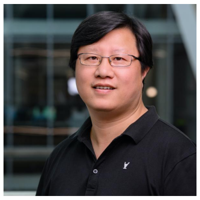
Zhuoxu Huang received a B.E. degree in geodesy and geomatics engineering from Wuhan University, Wuhan, China, in 2021. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Aberystwyth University, Aberystwyth, U.K. His research interests include video understanding, 3D vision, and machine learning.



Zhixiang Chen is a Lecturer in Machine Learning at the Department of Computer Science of the University of Sheffield. Before joining the University of Sheffield, he held postdoctoral positions at Imperial College London, UK and Tsinghua University, China. He received a PhD degree from Tsinghua University, China and a B.Eng. degree from Xi’an Jiaotong University, China.



Tao Xu received PhD degree from University of SurreyUK. He took several senior roles in AI industry and is currently the R&D director of Shanghai Investigation Design & Research Institute. His research interests include audio processing, computer vision, machine learning and multimodal large language models.



Jungong Han is Chair Professor in Computer Vision at the Department of Computer Science, the University of Sheffield, UK. He also holds an Honorary Professorship at the University of Warwick, UK. Previously, he was Chair Professor and Director of Research of the Computer Science department with Aberystwyth University, UK; Data Science Associate Professor with the University of Warwick; and Senior Lecturer in Computer Science with Lancaster University, UK.

Josef Kittler received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is a distinguished Professor of Machine Intelli gence at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and im age database retrieval, medical image analysis, and cognitive vision. He published the textbook Pattern Recognition: A Statistical Approach and over 700 scientific papers. His publications have been cited around 70,000 times (Google Scholar). He is series editor of Springer Lecture Notes on Computer Science. He currently serves on the Editorial Boards of Pattern Recognition Letters, Pattern Recognition and Artificial Intelligence, Pattern Analysis and Applications. He also served as a member of the Editorial Board of IEEE Transactions on Pattern Analysis and Machine Intelligence during 1982-1985. He served on the Governing Board of the International Association for Pattern Recognition (IAPR) as one of the two British representatives during the period 1982-2005, President of the IAPR during 1994-1996.